Entity-Based Tracking

David Eichmann

The University of Iowa
School of Library and Information Science

talk at http://mingo.info-science.uiowa.edu/eichmann/tdt2000.pdf

Hypothesis for TREC-9 / TDT2000 Work

- A focus on entities rather than words can yield superior performance
- Caveat: this is probably more valid for English only than for multilingual...
- But what the heck, let’s try!

Lexical Architecture

- Primary lexical scanner is custom written for TDT-style document formats
  - Dictionary-driven phrase recognition a clickable option
    - WordNet
    - Moby database
    - local instance generated from bibliographic citation keywords
- Alternative lexers implement
  - Wu’s Mandarin segmenter
  - Peterson’s Mandarin segmenter
  - Brill’s rule-driven POS tagger

Lexical Architecture, con’t.

- Lexical analysis is now supported as a cascade of filters
  - Initial token acquisition (including mapping encodings to Unicode)
  - Word segmentation (when necessary)
  - Language transformation (optional)
  - Part-of-speech tagging (optional)
  - Entity extraction (optional)

Named Entity Recognition

- We have five categories currently being recognized
  - Persons
  - Organizations
  - Locations
  - Event (preliminary)
  - MeSH
- All categories are driven through examination of noun phrases recognized by the POS tagger (with special handling of certain glue words: ‘and,’ ‘of,’ ‘the,’ etc.)
- Named entity vectors are maintained separately from the term vector, weighted by their length and the frequency of the constituent terms

Person Recognition Resources

- Various Web lists of cultural names
  - Anglo, Chinese, Arab, Hebrew, Hindi, Indian, Japanese, Latino, Muslim, Russian
  - World leaders
- This is enriched with a set of pattern expressions for other instances
  - “President” <proper name>
  - <proper name> “III”

Organization & Event Recognition Resources

❑ International political organizations (from CIA Fact Book)

❑ Fortune 500 company list

❑ Global 500 company list

❑ This is enriched with a set of pattern expressions for other instances

  ❍ <proper name> “Incorporated”

  ❍ <proper name> “&” “Sons”

Location Recognition Resources

❑ We mine the text of the CIA Fact Book for variants of country names, administrative divisions, capitals, harbors, etc.

❑ Various Web lists of

  ❍ World cities

  ❍ U.S. Cities

  ❍ Rivers

  ❍ Lakes

❑ This is enriched with a set of pattern expressions for other instances

  ❍ <proper name> “Street”

  ❍ “Mount” <proper name>

MeSH Recognition Resources

- We first load and reconstruct the MeSH term tree
- We then load the concept descriptors, binding them into the tree and adding the synonyms
- Finally the supplements are added to support drugs and compounds

Mandarin Entity Recognition

- Handled separately using a scheme derived from Peterson
  - Regular expression matcher where atoms are words
- The rationale was to recognize entity phrases in the original language and then map entities to entities, rather than words to words
- While plausible in theory, it blew our scores due to very low translation rates
  - A “Chinese CIA Fact Book” would be really nice...

A Topic Example: 30001

- Persons
  - Hun Sen (1)
- Organizations
- Places
  - Cambodia (4)
- Events
- MeSH Terms
- C_{track}(norm) = 0.86

A Second Topic Example: 31029

- Persons
  - Geidar Aliev (4)
- Organizations
  - National Independence Party (2)
  - Baku State University (1)
- Places
  - capital Baku (1)
  - Baku (1)
  - Azerbaijan (1)
  - Soviet Union (1)
- Events
  - October (1)
- MeSH Terms
- C_{track}(norm) = 0.00 (2 Corr. Det., 0 Missed Stories, 0 False Alarms)

Newswire Entity Recognition Sample #1

- APW19981001.0262 [Israel(0.271), Jonathan Pollard(0.153), Benjamin Netanyahu(0.102), Bill Clinton(0.102), United States(0.055), ...]
- Persons
  - Bill Clinton (3)
  - Jonathan Pollard (8)
  - Moshe Fogel (2)
  - Benjamin Netanyahu (2)
  - Esther (1)
  - Israeli Embassy (1)
- Organizations
  - Cabinet (1)
- Places
  - Israel (16)
  - United States (5)
  - Washington (2)

Newswire Entity Recognition Sample #2

- APW19981001.0303 [Vladimir Meciar(0.119), Slovak Democratic Coalition(0.065), Slovakia(0.043), United States and Germany(0.043), ...]
- Persons
  - Vladimir Meciar (8)
  - Jozef Moravcik (2)
  - God (1)
  - Kalman Petocz (2)
- Organizations
  - Slovak Democratic Coalition (2)
  - Organization (1)
  - United States and Germany (1)
  - NATO (1)
  - European Union (1)
  - Hungarian Coalition Party (1)
- Places
  - Slovakia (4)
  - Europe (1)

MeSH Entity Recognition Sample #1

- Document: 89316080 - Multiple and repetitive uses of the extended hamstring V-Y myocutaneous flap .
    - An extended hamstring V-Y myocutaneous advancement flap is described that may be used to cover unusually large defects in the ischial region. Technical points that allow a large amount of flap advancement are discussed. Because of its large size, the flap can be raised and used on repeated occasions to repair defects from recurrent ischial pressure sores. Two patients are presented in whom the same flap was used repeatedly on multiple occasions, demonstrating the potential for preservation of future options in such patients when this flap is used.

Provided MeSH Keywords

- Case Report
- Decubitus Ulcer/SU
    - C17.800.893.289
- Human
- Male
- Methods
    - E05.581
    - H01.770.370
- Middle Age
    - M01.060.116.630
- Reoperation
    - E04.690
- Surgical Flaps/*
    - A10.850.710
    - E07.862.710
- Thigh
    - A01.378.592.867

Phrases Generated by Tagger

- [Multiple, and, repetitive, uses, of, the] ⇒ [Multiple][repetitive, uses]
- [hamstring, V-Y, myocutaneous, flap]
- [hamstring, V-Y, myocutaneous, advancement, flap]
- [large, defects]
- [ischial, region]
- [Technical, points]
- [large, amount]
- [flap, advancement]
- [large, size]
- [flap]
- [and]
- [occasions]
- [defects]
- [recurrent, ischial, pressure, sores]
- [patients]
- [same, flap]
- [multiple, occasions]
- [potential]
- [preservation]
- [future, options]
- [such, patients]
- [flap]

Entity Matching

- MeSH Terms
  - Surgical Flaps (6)
    - A10.850.710
    - E07.862.710
  - Decubitus Ulcer (1)
    - C17.800.893.289
  - Patients (2)
    - M01.643
  - Forecasting (1)
    - I01.320
- Other Phrases
  - flap advancement (1)
  - future options (1)
  - hamstring V-Y myocutaneous advancement flap (1)
  - hamstring V-Y myocutaneous flap (1)
  - ischial pressure sores (1)
  - ischial region (1)
  - repetitive uses (1)

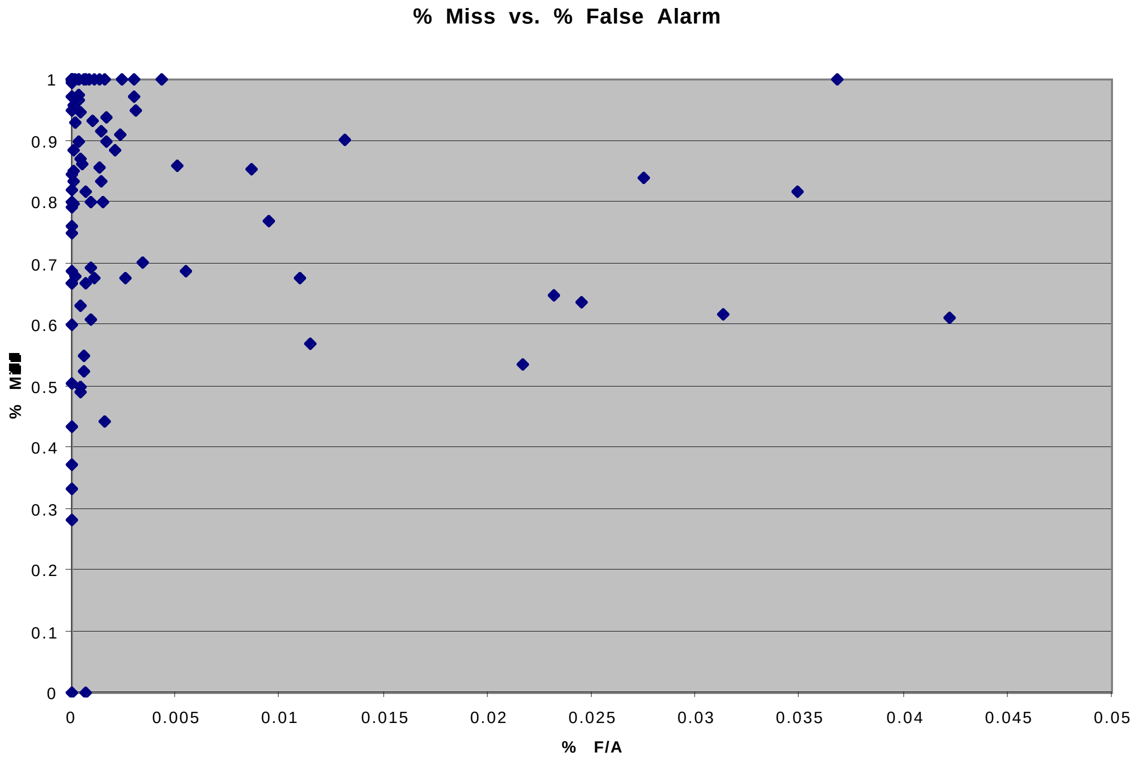
A Brief Record, MeSH Sample #2

- Document: 89316090 - Reconstructive surgery in Nicaragua [letter ; comment]
- Provided MeSH Keywords
  - Human
  - Nicaragua
    - Z01.107.169.690
  - Surgery, Plastic/*
    - G02.403.810.788
- Phrases
  - [Reconstructive, surgery]
  - [Nicaragua]
  - [letter]
- MeSH Terms
  - Surgery (1)
    - G02.403.810.762
  - Letter [Publication Type] (1)
- Other Phrases
  - Reconstructive surgery (1)

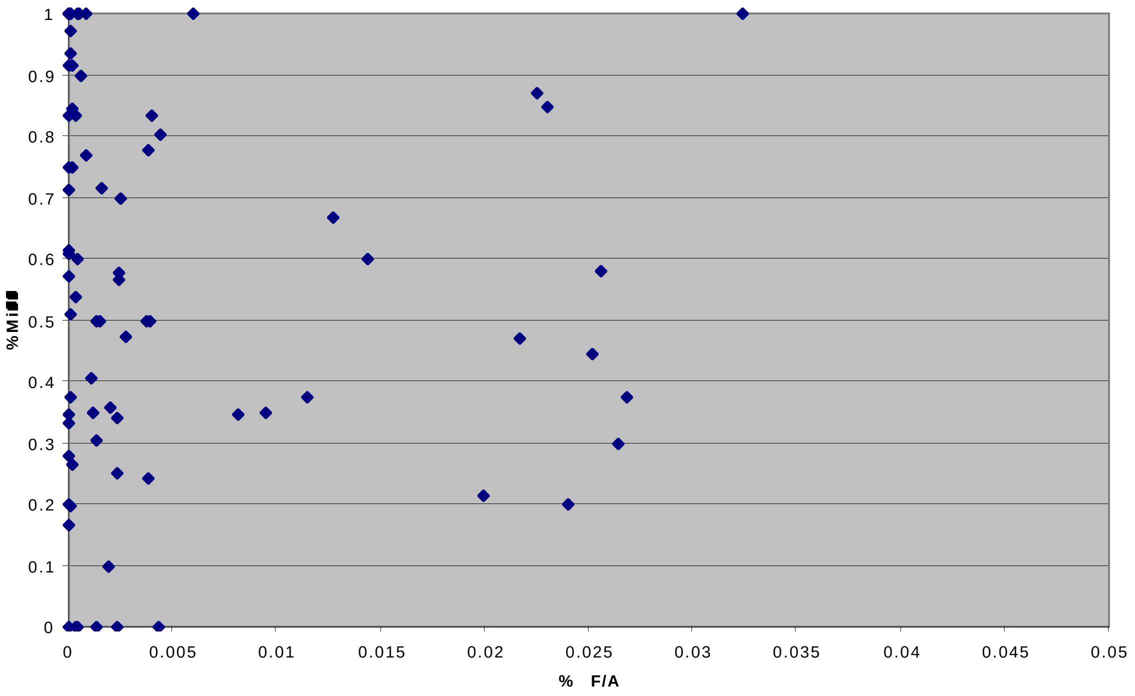
Composite Entity Similarity

- compute a cosine-vector score across each entity vector separately
- generate a weighted sum of the resulting similarities
    - 0.3 * sim(persons)
    - 0.3 * sim(organizations)
    - 0.2 * sim(locations)
    - 0.1 * sim(events)
    - 0.1 * sim(MeSH)

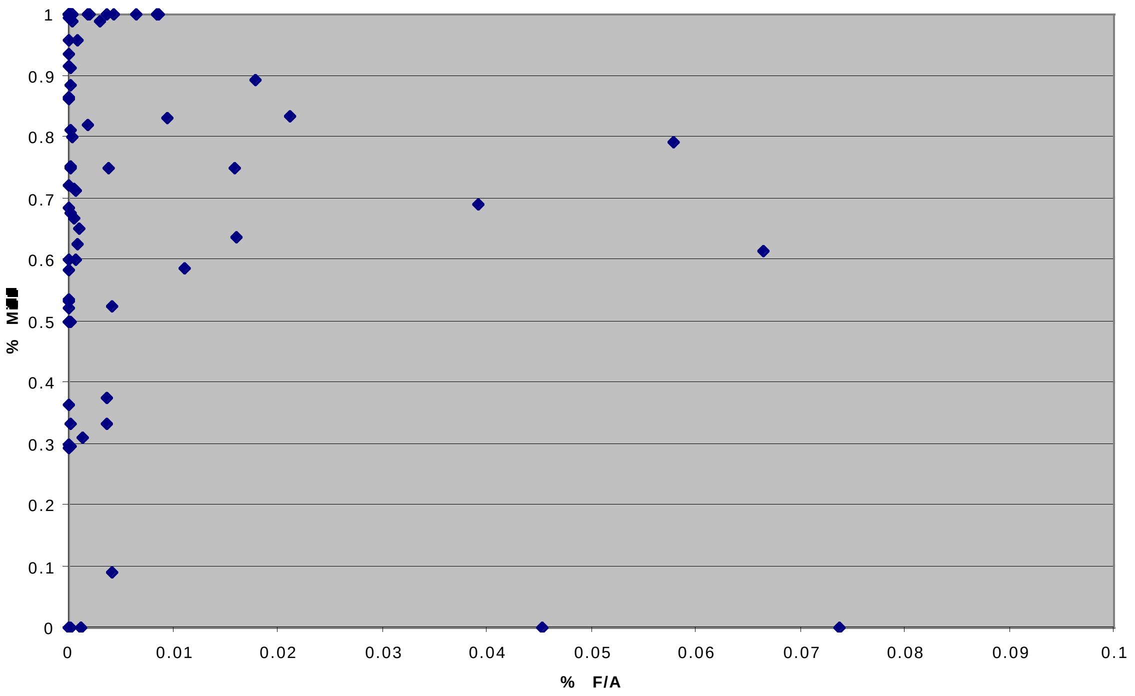
The DET Curve



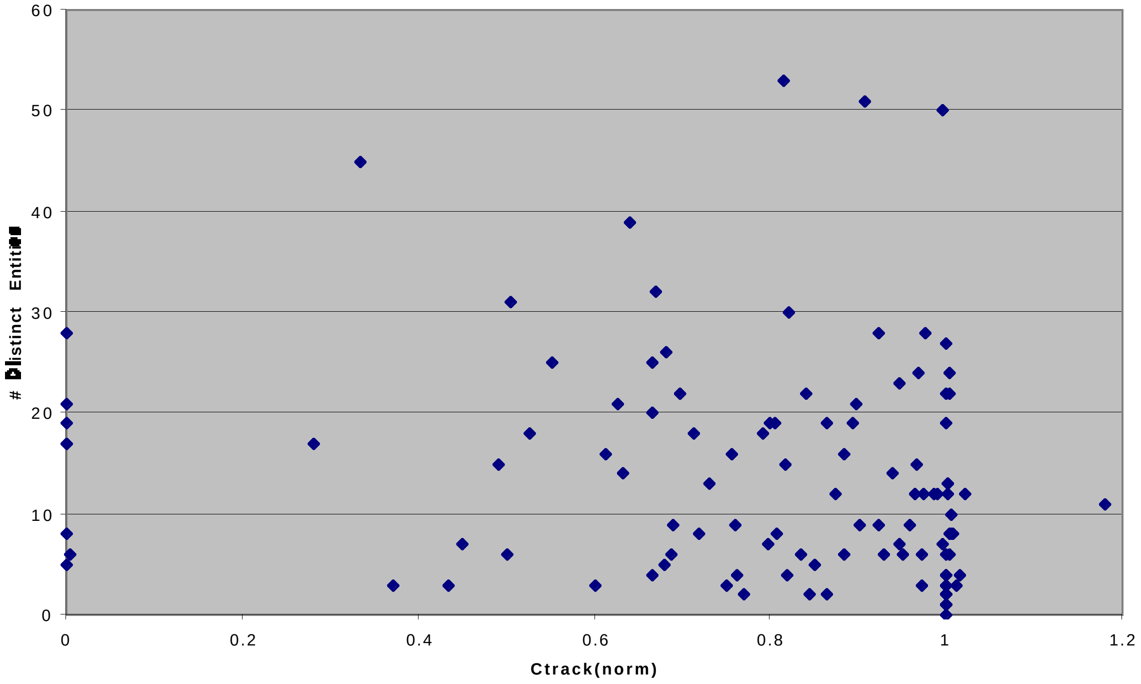
% Miss vs. % False Alarm (NWT English)



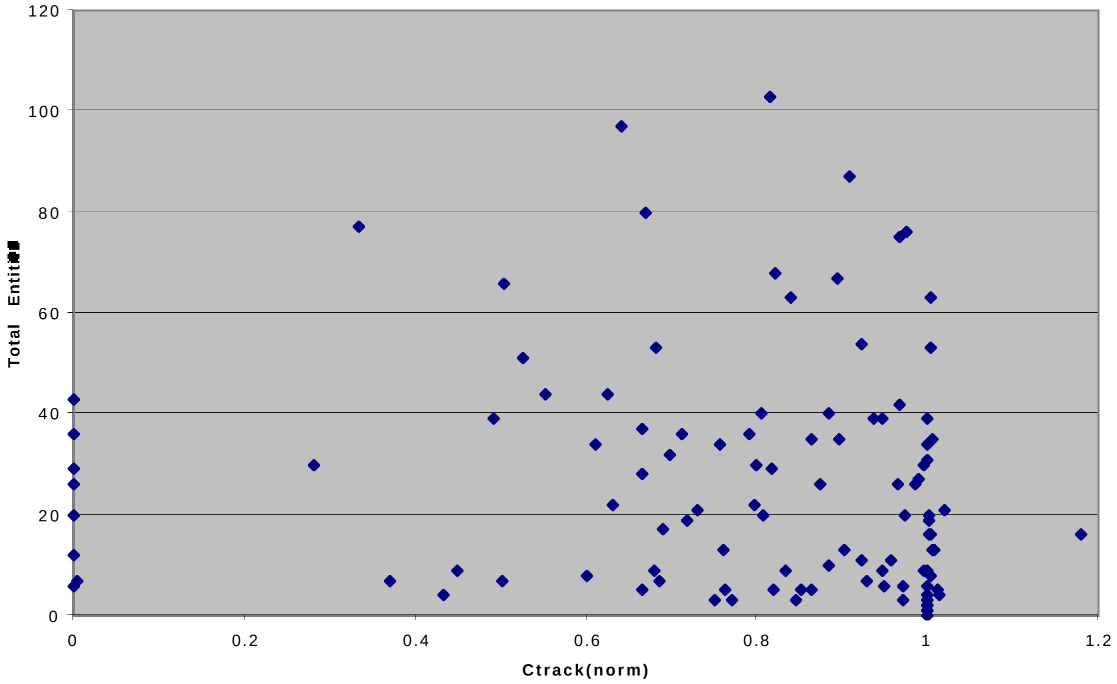
% Miss vs. % False Alarm (BN English)



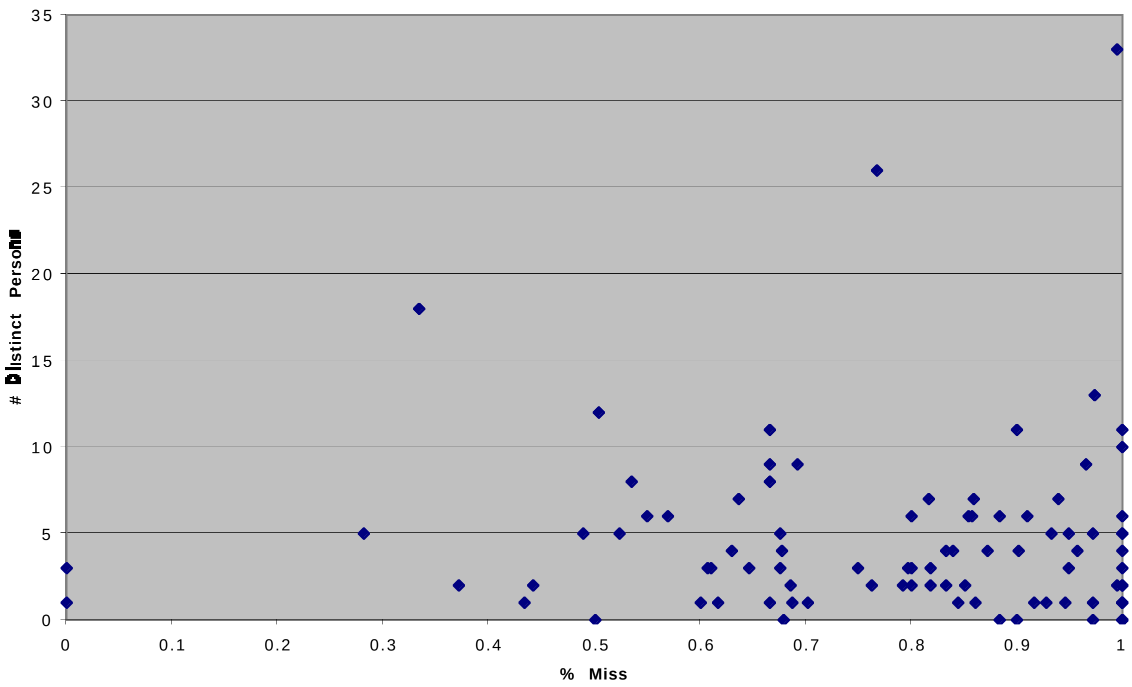
## Distinct Entities vs. Normalized Cost



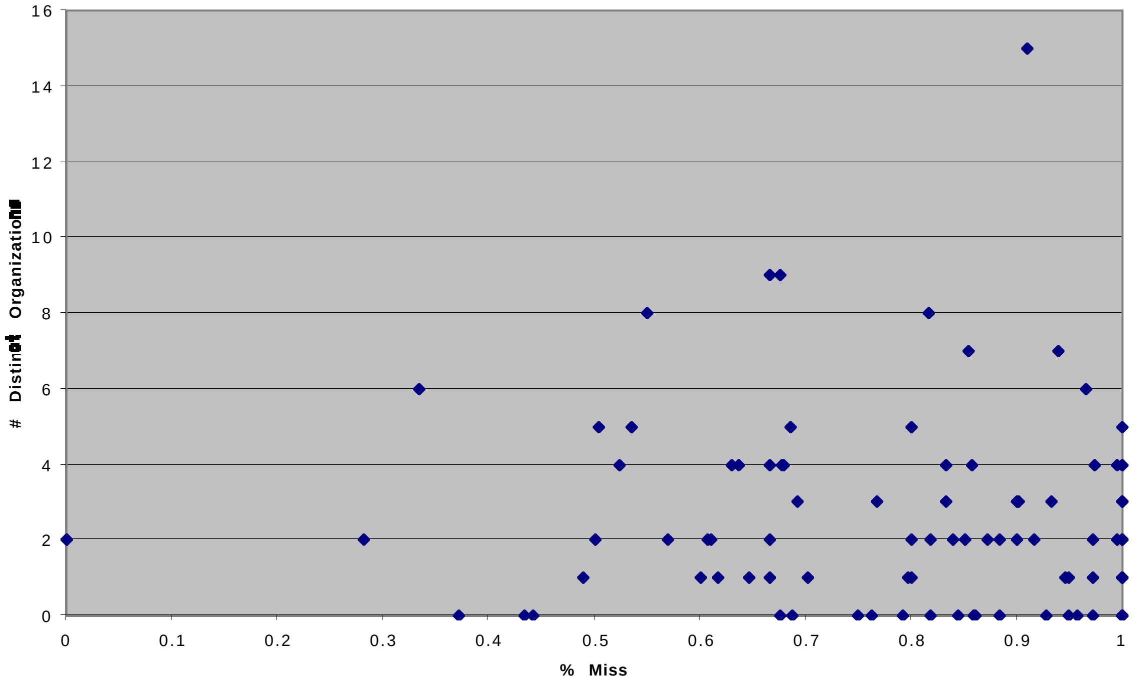
Total Entities vs. Normalized Cost



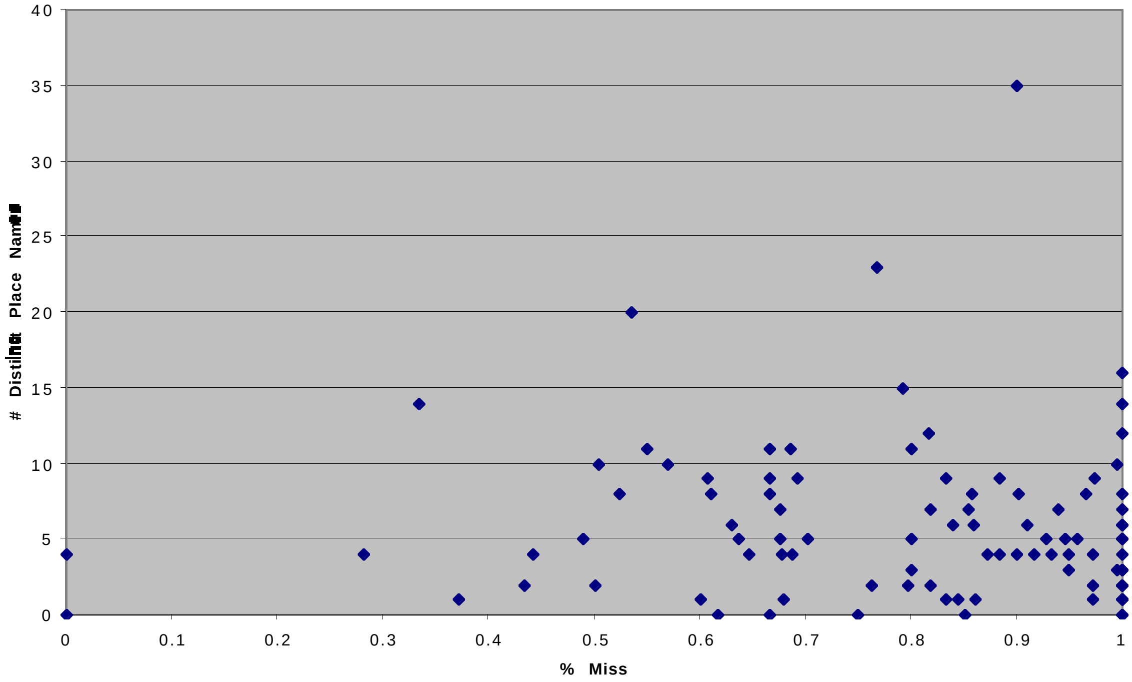
Distinct Persons vs. % Miss



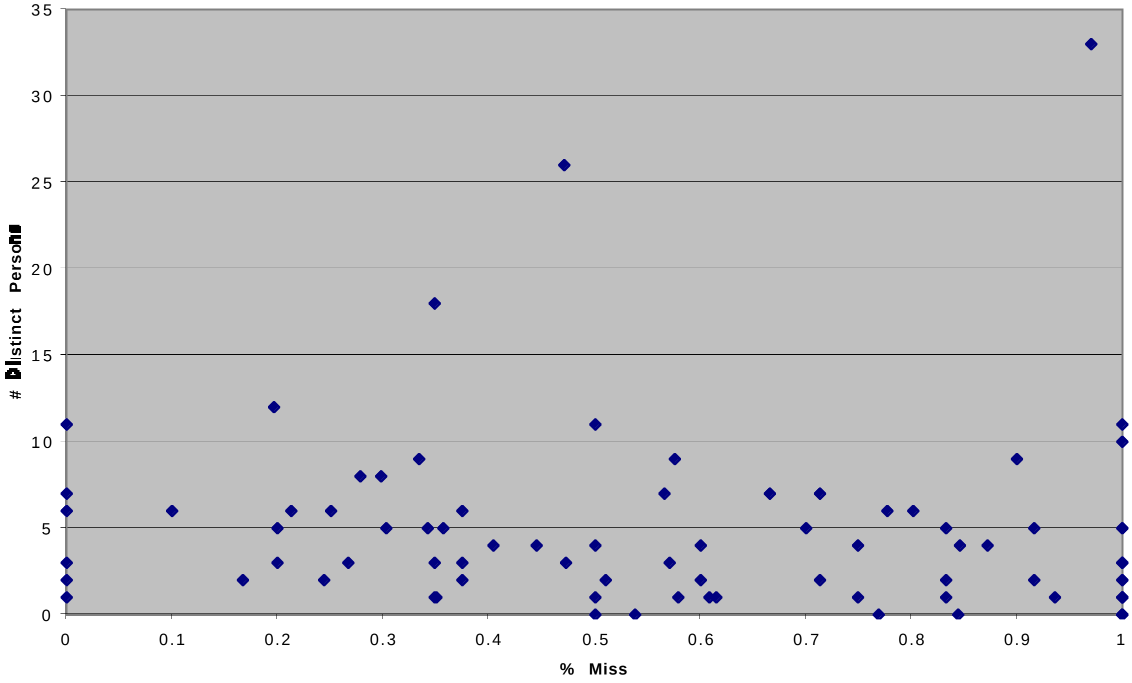
## Distinct Organizations vs. % Miss



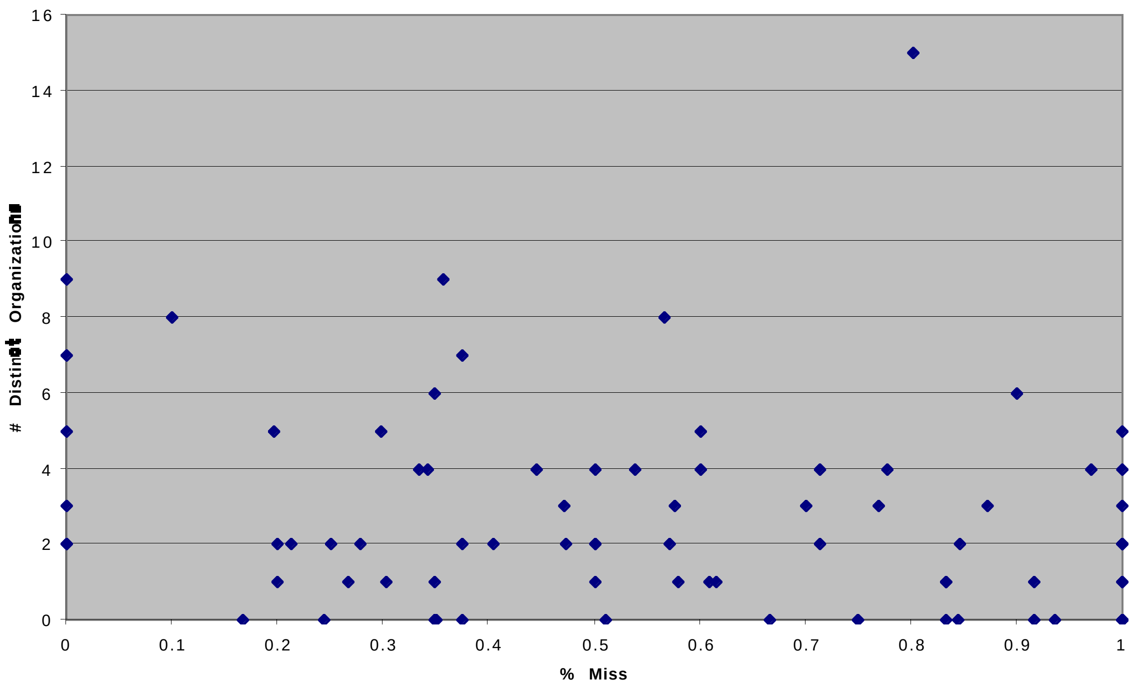
## Distinct Place Names vs. % Miss



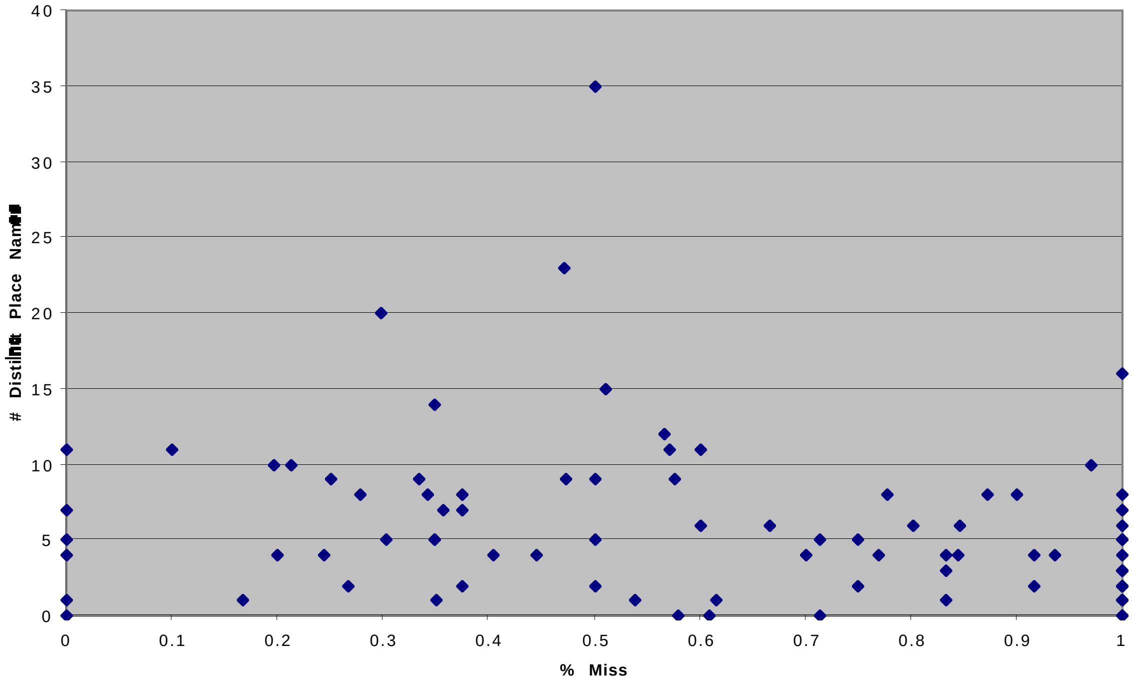
Distinct Persons vs. % Miss (NWT English)



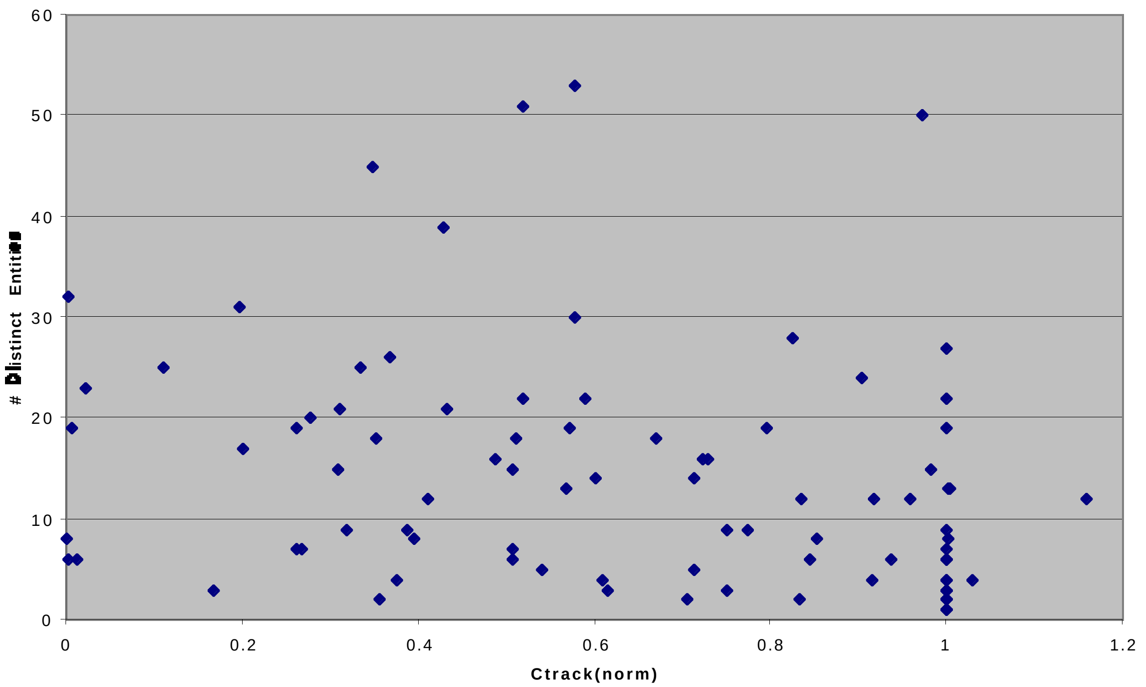
## Distinct Organizations vs. % Miss (NWT English)



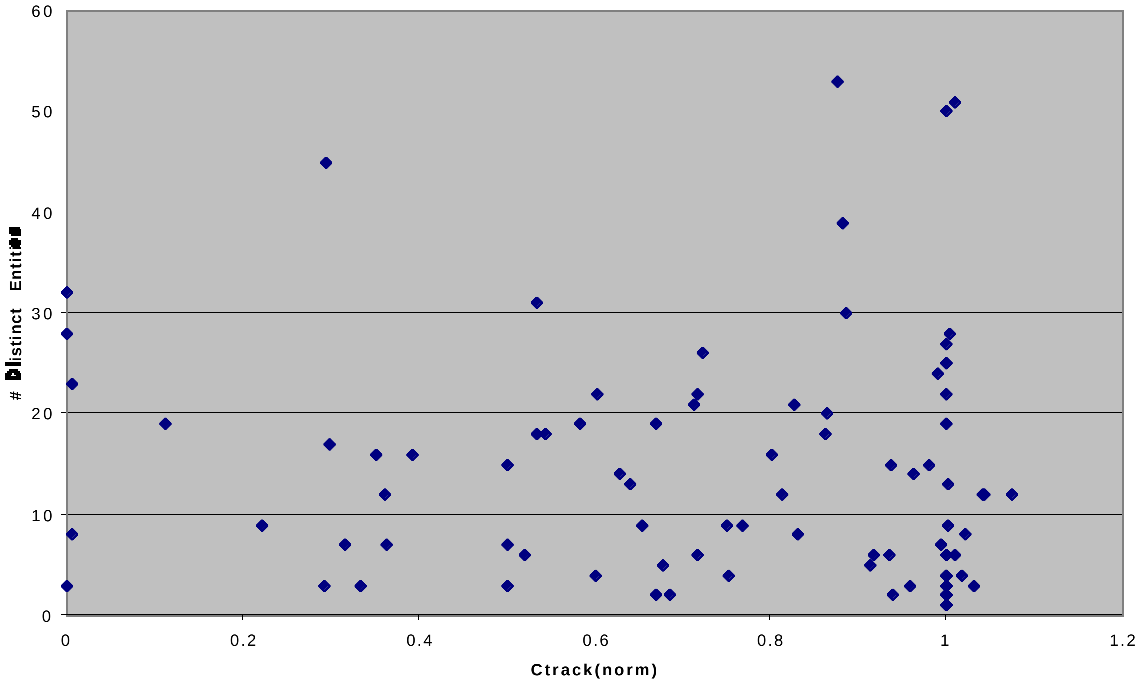
## Distinct Place Names vs. % Miss (NWT English)



## Distinct Entities vs. Normalized Cost (NWT English)



## Distinct Entities vs. Normalized Cost (BN English)



Multilingual Contrast Run

		Raw Counts			Scores		
		Corr. Det	Missed	F/A	P(Miss)	P(Fa)	C_{track}(norm)
With Entities	sum	2713	9345	15944	.8165	.0032	.8322
	mean	18	81	138			
Without Entities	sum	3739	8334	4752	.6289	.0010	.6337
	mean	32	72	41			

Multilingual Contrast Run (NWT English)

		Raw Counts			Scores		
		Corr. Det	Missed	F/A	P(Miss)	P(Fa)	C_{track}(norm)
With Entities	sum	1198	2230	5504	.6401	.0039	.6592
	mean	10	19	47			
Without Entities	sum	1656	1926	1990	.4487	.0014	.4555
	mean	14	16	17			

Conclusions

- Using a pure entity scheme shows potential
  - There are obvious impacts relating
    - coverage of entities domains
    - recognition levels
- Running a TREC Adaptive Filtering tuned system against the TDT Tracking task is viable if you’re more interested in precision than recall
- Whither TDTeval?
- talk at http://mingo.info-science.uiowa.edu/eichmann/tdt2000.pdf